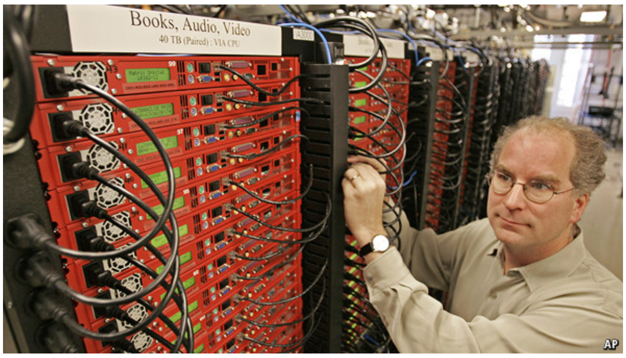
The Economist

Babbage
Science and technology

Internet permanence

Down the memory hole

Jan 15th 2014, 10:50 by G.F. | SEATTLE



THE Internet Archive, founded by Brewster Kahle, has the modest goal of maintaining a frequently updated copy of everything on the internet. It has many purposes (https://archive.org/about/) for its data, which include providing a contemporary record for researchers and a permanent one for historians. (Technically, Alexa (http://www.alexa.com/) , a for-profit firm founded by Kahle and sold in 1999 to Amazon, collects the data, but is under permanent contract to ship it after a six-month lag to the nonprofit archive.)

However, the archive's kryptonite is a simple file: robots.txt (https://archive.org/about/exclude.php) . This file's name refers to the automated "crawlers" run by AltaVista in the web's distant past (the late 1990s) and Google, Microsoft and many others today. Crawler methodically follow every link on a webpage, retrieve the contents and index them.

The robots.txt file is a decidedly low-tech way to tell all robots or specific ones to not index or archive a site. It's just a bit of text with a format that identifies robots by a short, agreed-upon name, and tells them what they may and may not do. Responsible search engines and archives honour its settings. There is essentially no enforcement; it is tacit.

This is why when one uses the archive's colourfully named Wayback Machine (https://archive.org/web/) to see the history of pages on a given website and no results appear or the results are limited, one knows the website in question has opted to drop previous versions of its pages down the memory hole rather than consign them to a public and permanent record.

No country has a mandate to record every webpage published within its borders; the Internet Archive is the closest thing in the world to such a comprehensive effort. (Some governments and government agencies have specific obligations and commitments.) And material can be deleted so easily and permanently as to distort the memory of public discourse.

On January 13th, the *Guardian* removed an essay (http://www.theguardian.com/commentisfree/2014/jan/08/lisa-adams-tweeting-cancer-ethics) by Emma Keller titled "Forget funeral selfies", and which dealt with Ms Keller's reaction and interaction with cancer patient Lisa Adams, who has used Twitter and other social networks to discuss her prognosis, care and progress exhaustively. After Ms Keller's husband, Bill, published a different essay in the *New York Times*, a social-network uproar on January 12th led the *Guardian* to remove the text, rendering the context of the discussion unavailable.

The *Guardian* left comments intact, but initially replaced the essay with the text, "This post has been deleted with the agreement of the subject because it is inconsistent with the Guardian editorial code." Later in the day, that text was itself deleted and replaced with "This post has been removed pending investigation."

Websites are of different minds as to how they label discredited, withdrawn or revised work. Some throw pieces in the bit bucket, and they are gone for good. Others annotate and leave the original up. This is particularly common with withdrawn research. The long-delayed retraction of the Andrew Wakefield autism study still appears (http://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2897%2911096-0/abstract) on the *Lancet*'s website, but with a large red banner reading "retracted", which is also found in the page's title and elsewhere. Providing a retraction alongside the original paper allows easier refutation later of its points, if raised.

The *New York Times* thoroughly documented and annotated (http://www.nytimes.com/ref/national/BLAIR-ARCHIVE.html) the journalistic sins of fabulist Jayson Blair on its site, providing a helpful summary of articles found wanting, a collation of the deficits and links to each article which are also marked up.

The internet is often seen as a memory palace rather than a memory hole. Many people find detritus of past lives wash up long after they thought a photo or fact was drowned. But the internet can also be an amnesiac: what we thought was indelible can be wiped out without leaving a trace.